Feature Grouping as a Stochastic Regularizer for High Dimensional Structured Data

Sergül Aydöre
(Stevens Institute of Technology, USA)

Bertrand Thirion
(INRIA, France)

Gaël Varoquaux
(INRIA, France)
High Dimensional and Small-Sample Data Situations

- Brain imaging, Genomics, Seismology, Astronomy, Chemistry, etc.

PET acquisition process wikipedia

MRI Scanner and rs-fMRI time series acquisition [NVIDIA]

A typical MEG equipment [BML2001]

Genomics Integrative Genomics Viewer, 2012

Seismology https://www.mapnagroup.com

Astronomy Astronomy Magazine, 2015
Fitting Complex Models in These Situations

Challenges

1. **Large feature dimension**: due to rich temporal and spatial resolution
2. **Noise in the data**: due to artifacts unrelated to the effect of interest
3. **Small sample size**: due to logistics and cost of data acquisition

Regularization Strategies

- **Early Stopping**: [Yao, 2007]
- **ℓ₁ and ℓ₂ penalties**: [Tibshirani 1996]
- **Pooling Layers in CNNs**: [Hinton 2012]
- **Group LASSO**: [Yuan 2006]
- **Dropout**: [Srivastana 2014]
Fitting Complex Models in These Situations

Challenges

1. Large feature dimension: due to rich temporal and spatial resolution
2. Noise in the data: due to artifacts unrelated to the effect of interest
3. Small sample size: due to logistics and cost of data acquisition

Regularization Strategies

- Early Stopping: [Yao, 2007]
- \(\ell_1\) and \(\ell_2\) penalties: [Tibshirani 1996]
- Pooling Layers in CNNs: [Hinton 2012]
- Group LASSO: [Yuan 2006]
- Dropout: [Srivastava 2014]
Fitting Complex Models in These Situations

**Challenges**

1. **Large feature dimension**: due to rich temporal and spatial resolution
2. **Noise in the data**: due to artifacts unrelated to the effect of interest
3. **Small sample size**: due to logistics and cost of data acquisition

**Regularization Strategies**

- **Early Stopping**: [Yao, 2007]
- **$\ell_1$ and $\ell_2$ penalties**: [Tibshirani 1996]
- **Pooling Layers in CNNs**: [Hinton 2012] ....................... TRANSLATION INVARINANCE
- **Group LASSO**: [Yuan 2006] ................................. STRUCTURE + SPARSITY
- **Dropout**: [Srivastana 2014] .................................. STOCHASTICITY

- **PROPOSED**: Use STRUCTURE & STOCHASTICITY

**POSTER**: Pacific Ballroom #121, 06/11, Tuesday
**Feature Grouping to Capture Structure**

**Algorithm**

**Training Data**

- **Recursive Nearest Agglomeration (ReNA)**
  - [Hoyos et al 2016]

- **Number of clusters = 5**

- **Iteration 1**: Initial assignment of features to clusters.
- **Iteration 2**: Clusters are merged.
- **Iteration N**: Clusters are recursively merged until the desired number of clusters remain.

- **ReNA**: a data-driven, graph constrained feature grouping algorithm

- Each feature (pixel) is assigned to a cluster. Clusters are then recursively merged until the desired number of clusters remain.

- Benefits of ReNA: 
  1. A fast clustering algorithm
  2. Leads to good signal approximations.

**Feature Grouping Matrix** $\Phi \in \mathbb{R}^{k \times p}$

$\Phi = \begin{bmatrix}
\alpha_1 & \cdots & \alpha_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & \cdots & 0 & \alpha_2 & \cdots & \alpha_2 & 0 & \cdots & 0 \\
0 & \cdots & 0 & 0 & \cdots & 0 & \alpha_3 & \cdots & \alpha_3 \\
0 & \cdots & 0 & \cdots & 0 & \cdots & \alpha_4 & \cdots & \alpha_4 \\
0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & \alpha_5 \\
\end{bmatrix}$

- Each row captures a different structure

**Reduction and Low-rank Approximation**

- $x \in \mathbb{R}^p$
- $\Phi x \in \mathbb{R}^{k \times p}$
- $k \ll p$
- $\Phi^T \Phi x \in \mathbb{R}^p$

**POSTER**: Pacific Ballroom #121, 06/11, Tuesday
Proposed Approach

Consider fully connected neural network with $H$ layers

Algorithm 1: Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate $\eta$

Require: Initial Parameters for $H$ layers

\[ \Theta \triangleq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\} \]

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling $r$ samples from the training data set with replacement

\[ \Phi = \{\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(l)}\} \]

1: while stopping criteria not met do

2: Sample a minibatch of $m$ samples from the training set \( \{x^{(1)}, \ldots, x^{(m)}\} \) with corresponding labels \( y^{(i)} \)

3: Sample \( \Phi \) from the bank \( \Phi \).

4: Define \( \Xi \triangleq \left\{ W_0, b_0, W_1, b_1, \ldots, W_H, b_H \right\} \) where $W_0 \triangleq W_0 \Phi^T$.

5: Compute gradient estimate:

\[ g \leftarrow \frac{1}{m} \nabla_{\Xi} \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right) \]

6: Apply updates:

- $W_0 \leftarrow W_0 - \eta g_{w_0}$
  where $g_{w_0} \triangleq \frac{1}{m} \nabla_{W_0} \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$

- $b_j \leftarrow b_j - \eta g_{b_j}$
  where $g_{b_j} \triangleq \frac{1}{m} \nabla_{b_j} \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$
  for $j \in \{0, \ldots, H\}$

- $W_j \leftarrow W_j - \eta g_{w_j}$
  where $g_{w_j} \triangleq \frac{1}{m} \nabla_{W_j} \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$
  for $j \in \{1, \ldots, H\}$

7: end while
Proposed Approach

Pre-compute a bank of feature grouping matrices

Algorithm 1: Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate $\eta$
Require: Initial Parameters for $H$ layers
\[ \Theta \triangleq \{ W_0, b_0, W_1, b_1, \ldots, W_H, b_H \} \]

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling $r$ samples from the training data set with replacement
\[ \Phi = \{ \Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(b)} \} \]

1: while stopping criteria not met do
2:   Sample a minibatch of $m$ samples from the training set \( \{ x^{(1)}, \ldots, x^{(m)} \} \) with corresponding labels \( y^{(i)} \)
3:   Sample \( \Phi \) from the bank \( \Phi \).
4:   Define \( \Xi \triangleq \left\{ W_0, b_0, W_1, b_1, \ldots, W_H, b_H \right\} \) where
   \[ W_0 \triangleq W_0 \Phi^T \]
5:   Compute gradient estimate:
   \[ g \leftarrow \frac{1}{m} \nabla_{\Xi} \sum_i L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right) \]
6:   Apply updates:
   - \( W_0 \leftarrow W_0 - \eta g_{w_0} \Phi \)
     where \( g_{w_0} \triangleq \frac{1}{m} \nabla_{W_0} \sum_i L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right) \)
   - \( b_j \leftarrow b_j - \eta g_{b_j} \)
     where \( g_{b_j} \triangleq \frac{1}{m} \nabla_{b_j} \sum_i L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right) \)
     for \( j \in \{0, \ldots, H\} \)
   - \( W_j \leftarrow W_j - \eta g_{w_j} \)
     where \( g_{w_j} \triangleq \frac{1}{m} \nabla_{W_j} \sum_i L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right) \)
     for \( j \in \{1, \ldots, H\} \)
7: end while
Proposed Approach

Sample from the training set

Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate $\eta$

Require: Initial Parameters for $H$ layers

$\Theta \triangleq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling $r$ samples from the training data set with replacement

$\Phi = \{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(b)}\}$

1: while stopping criteria not met do
2: Sample a minibatch of $m$ samples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$ with corresponding labels $y^{(i)}$
3: Sample $\Phi$ from the bank $\Phi$.
4: Define $\Xi \triangleq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$ where $W_0 \triangleq W_0 \Phi^T$.
5: Compute gradient estimate:

$g \leftarrow \frac{1}{m} \nabla_{\Xi} \sum_{i} L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$

6: Apply updates:

- $W_0 \leftarrow W_0 - \eta g_{w_0}$ where $g_{w_0} \triangleq \frac{1}{m} \nabla_{W_0} \sum_{i} L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$

- $b_j \leftarrow b_j - \eta g_{b_j}$ where $g_{b_j} \triangleq \frac{1}{m} \nabla_{b_j} \sum_{i} L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$ for $j \in \{0, \ldots, H\}$

- $W_j \leftarrow W_j - \eta g_{w_j}$ where $g_{w_j} \triangleq \frac{1}{m} \nabla_{W_j} \sum_{i} L \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$ for $j \in \{1, \ldots, H\}$

7: end while
Proposed Approach

Sample $\Phi$ from the bank of feature grouping matrices

Algorithm 1: Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate $\eta$
Require: Initial Parameters for $H$ layers

$\Theta \triangleq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling $r$ samples from the training data set with replacement

$\Phi = \{\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(b)}\}$

1: while stopping criteria not met do
2: Sample a minibatch of $m$ samples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$ with corresponding labels $y^{(i)}$
3: Sample $\Phi$ from the bank $\Phi$.
4: Define $\Xi \triangleq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$ where $W_0 \triangleq W_0 \Phi^T$.
5: Compute gradient estimate:

$$g \leftarrow \frac{1}{m} \nabla \Xi \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$$
6: Apply updates:

- $W_0 \leftarrow W_0 - \eta g_{w_0}$
  where $g_{w_0} \triangleq \frac{1}{m} \nabla W_0 \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$
- $b_j \leftarrow b_j - \eta g_{b_j}$
  where $g_{b_j} \triangleq \frac{1}{m} \nabla b_j \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$ for $j \in \{0, \ldots, H\}$
- $W_j \leftarrow W_j - \eta g_{w_j}$
  where $g_{w_j} \triangleq \frac{1}{m} \nabla W_j \sum_i \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$ for $j \in \{1, \ldots, H\}$
7: end while
Proposed Approach

Re-define parameter space and project input onto lower dimensional space

Algorithm 1: Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate $\eta$
Require: Initial Parameters for $H$ layers

$\Theta \doteq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling $r$ samples from the training data set with replacement

$\Phi = \{\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(b)}\}$

1: while stopping criteria not met do
2: Sample a minibatch of $m$ samples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$ with corresponding labels $y^{(i)}$
3: Sample $\Phi$ from the bank $\Phi$.
4: Define $\Xi \doteq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$ where $W_0 \doteq W_0 \Phi^T$
5: Compute gradient estimate:

$g \leftarrow \frac{1}{m} \nabla_{\Xi} \sum_{i} L \left(f(\Phi x^{(i)}; \Xi), y^{(i)}\right)$

6: Update parameters:

- $W_0 \leftarrow W_0 - \eta g_{w_0, \Phi}$
  where $g_{w_0} \doteq \frac{1}{m} \nabla_{w_0} \sum_{i} L \left(f(\Phi x^{(i)}; \Xi), y^{(i)}\right)$
- $b_j \leftarrow b_j - \eta g_{b_j}$
  where $g_{b_j} \doteq \frac{1}{m} \nabla_{b_j} \sum_{i} L \left(f(\Phi x^{(i)}; \Xi), y^{(i)}\right)$
  for $j \in \{0, \ldots, H\}$
- $W_j \leftarrow W_j - \eta g_{w_j}$
  where $g_{w_j} \doteq \frac{1}{m} \nabla_{w_j} \sum_{i} L \left(f(\Phi x^{(i)}; \Xi), y^{(i)}\right)$
  for $j \in \{1, \ldots, H\}$

7: end while
Proposed Approach

Apply back propagation

Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate $\eta$
Require: Initial Parameters for $H$ layers

$\Theta \triangleq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling $r$ samples from the training data set with replacement

$\Phi = \{\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(b)}\}$

1: while stopping criteria not met do
2: Sample a minibatch of $m$ samples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$ with corresponding labels $y^{(i)}$
3: Sample $\Phi$ from the bank $\Phi$
4: Define $\Xi \triangleq \{W_0, b_0, W_1, b_1, \ldots, W_H, b_H\}$ where $W_0 \triangleq W_0 \Phi^T$
5: Compute gradient estimate:

$g \leftarrow \frac{1}{m} \nabla_{\Xi} \sum_{i} \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$

6: Apply updates:

- $W_0 \leftarrow W_0 - \eta g_{w_0}$, where $g_{w_0} \triangleq \frac{1}{m} \nabla_{W_0} \sum_{i} \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$
- $b_j \leftarrow b_j - \eta g_{b_j}$, where $g_{b_j} \triangleq \frac{1}{m} \nabla_{b_j} \sum_{i} \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$ for $j \in \{0, \ldots, H\}$
- $W_j \leftarrow W_j - \eta g_{w_j}$, where $g_{w_j} \triangleq \frac{1}{m} \nabla_{W_j} \sum_{i} \mathcal{L} \left( f(\Phi x^{(i)}; \Xi), y^{(i)} \right)$ for $j \in \{1, \ldots, H\}$
7: end while
Proposed Approach

Update parameters

To update $W_0$, project gradients back to the original space.

Other terms are updated in a standard way.
Experimental Results

Noisy Settings

Performance in terms of computation time for Olivetti Faces

Performance in terms of sample size for fMRI data

Small-sample Settings

Feature Grouping performs best as the sample size decreases

Feature Grouping is computationally efficient and robust to noise
Thank You!

Visit our POSTER TODAY at Pacific Ballroom #121!